UNITED STATES PATENT APPLICATION

FOR

Detection of Characteristics of Human-Machine Interactions for Dialog Customization and Analysis

INVENTORS:

Michael H. Cohen
Larry P. Heck
Jennifer E. Balogh
James M. Riseman
Naghmeh N. Mirghafori

Prepared by:

BLAKELY, SOKOLOFF, TAYLOR & ZAFMAN LLP
12400 WILSHIRE BOULEVARD
SEVENTH FLOOR
LOS ANGELES, CALIFORNIA 90025
(408) 720-8300

Attorney's Docket No. 3932P006XX

Detection of Characteristics of Human-Machine Interactions for Dialog Customization and Analysis

[0001] This is a continuation-in-part of U.S. Patent application no. 09/412,173, filed on October 4, 1999 and entitled, "Method and Apparatus for Optimizing a Spoken Dialog Between a Person and a Machine", which is a continuation-in-part of U.S. Patent application no. 09/203,155, filed on December 1, 1998 and entitled, "System and Method for Browsing a Voice Web", each of which is incorporated herein by reference in its entirety.

FIELD OF THE INVENTION

[0002] The present invention pertains to systems which use automatic speech recognition. More particularly, the present invention relates to detection of characteristics of human-machine interactions for purposes of dialog customization and analysis.

BACKGROUND OF THE INVENTION

[0003] Speech applications are rapidly becoming commonplace in everyday life. A speech application may be defined as a machine-implemented application that performs tasks automatically in response to speech of a human user and which responds to the user with audible prompts, typically in the form of recorded or synthesized speech. For example, speech applications may be designed to allow a user to make travel reservations or buy stock over the telephone, without assistance from a human operator. The interaction between the person and the machine is referred to as a dialog.

[0004]   Automatic speech recognition (ASR) is a technology used to allow machines to recognize human speech. Commonly, an ASR system includes a speech recognition engine, which uses various types of data models to recognize an utterance. These models typically include language models, acoustic models, grammars, and a dictionary.

[0005]   It is desirable for speech applications and speech recognition systems to provide more personalized experiences for their users and to respond more intelligently to the users and their environments. In addition, it is desirable to have the ability to analyze accumulated data representing dialogs, to identify demographics and other characteristics of the users and their environments.

2

SUMMARY OF THE INVENTION

[0006] According to one aspect of the present invention, an audio-based dialog between a person and a machine is established, wherein the person uses a communication device to communicate with the machine. A characteristic is automatically detected during the dialog in real time, wherein the characteristic is not uniquely indicative of any of: the identity of the person, the identity of the specific communication device, or any user account. The dialog is then customized at an application level, based on the detected characteristic.

[0007] According to another aspect of the present invention, multiple audio-based dialogs are provided, each between a person and a machine, wherein each person uses a communication device to communicate with the machine during the corresponding dialog. Each of the dialogs is examined to automatically detect a characteristic for at least some of the dialogs, wherein the characteristic is not uniquely indicative of any of: the identity of the person, the identity of the specific communication device, or any user account. An overall characterization of the dialogs is then generated with respect to the characteristic.

[0008] The present invention further includes an apparatus corresponding to each of these methods.

[0009] Other features of the present invention will be apparent from the accompanying drawings and from the detailed description which follows.

3

BRIEF DESCRIPTION OF THE DRAWINGS

[0010]   The present invention is illustrated by way of example and not

limitation in the figures of the accompanying drawings, in which like references

indicate similar elements and in which:

[0011]   Figure 1 illustrates a network configuration in which a person

communicates with a speech recognition system;

[0012]   Figure 2 illustrates a process of customizing dialog based on detected

characteristics, according to one embodiment;

[0013]   Figure 3 illustrates an embodiment of the speech recognition system;

[0014]   Figure 4 illustrates two embodiments of a process for analyzing dialog

data based on detected characteristics;

[0015]   Figure 5 illustrates a process for training a statistical classifier;

[0016]   Figure 6 illustrates a process of using training classification-specific

acoustic models to make classification decisions;

[0017]   Figure 7 illustrates a process of applying multiple simultaneous

classifications to make a classification decision, using caller-independent

transforms; and

[0018]   Figure 8 is a high-level block diagram of a processing system in which

the speech recognition system can be implemented.

DETAILED DESCRIPTION

[0019] Described below are a method and apparatus for customizing a human-machine dialog or analyzing data relating to human-machine dialogs, based on detection of characteristics associated with one or more dialogs. Note that in this description, references to "one embodiment" or "an embodiment" mean that the feature being referred to is included in at least one embodiment of the present invention. Further, separate references to "one embodiment" in this description do not necessarily refer to the same embodiment; however, neither are such embodiments mutually exclusive, unless so stated, and except as will be readily apparent to those skilled in the art. For example, a feature, structure, act, etc. described in one embodiment may also be included in other embodiments. Thus, the present invention can include a variety of combinations and/or integrations of the embodiments described herein.

I. Overview

[0020] As described in greater detail below, a speech recognition system automatically detects one or more characteristics associated with a dialog and uses the detected characteristic(s) to customize (personalize) the dialog for the human speaker in real-time at the application level. A customized dialog provides the speaker with a richer, more efficient and more enjoyable experience. Customizing a dialog "at the application level" may involve, for example, customizing features such as call routing, error recovery, call flow, content selection, system prompts, or the speech recognition system's persona.

5

[0021] In addition (or as an alternative) to customizing a dialog, detected characteristics may be used to generate an offline analysis of dialog data. Off-line analysis of dialog data based on detected characteristics can be used for many purposes, such as to enable marketing professionals to understand the demographics of their callers (e.g., age, gender, accent, etc.). These demographics can be used to improve service, enhance targeted advertising, etc.

[0022] The characteristics that are automatically detected by the system may include, for example, characteristics of the speaker, his speech, his environment, or the speech channel (e.g., the type of communication device used by the speaker to communicate with the speech recognition system). The system described herein is not specifically directed to detecting characteristics that uniquely identify a speaker, his communication device or user account (e.g., caller ID information, automatic number identification (ANI), device IDs, user IDs, and the like); nonetheless, embodiments of the system may detect such information in addition to detecting the characteristics mentioned below.

[0023] The system described herein improves the user's experience by detecting one or more of a broad range of characteristics, some of which are quite subtle and difficult to detect without directly asking the speaker. Thus, the system automatically detects various personalizing statistical classifications and direct measurements of signal characteristics. The statistical classifications can include information about the speaker and/or the acoustics associated with the immediate environment of the speaker and the current communication link (e.g.,

telephone connection). The detected speaker characteristics may be, for example: the gender of the speaker, the type of speech being spoken by the speaker (e.g., side speech, background speech, fast speech, slow speech, accented speech), the emotional state of the speaker, whether the speaker is telling the truth, the approximate age of the speaker, the physical orientation of the speaker (e.g. standing, sitting, walking), and/or the apparent health of the speaker (e.g. sick, congested, etc.).

[0024] The acoustic characteristics that are automatically detected by the system may include, for example: the channel type (e.g., hands-free mobile telephone vs. hand-held mobile telephone, type of codec, quality of transmission, microphone type), classifications of the speaker's acoustic environment (e.g., in a car, in a car on a highway, in an airport or crowded hall, at a restaurant or noisy bar, on a sidewalk of a busy street, at home, at work).

[0025] The direct measurements of the signal characteristics may include, for example, the level (e.g., energy) of the speech, the level of the background noise (described on a dB scale), and the signal-to-noise ratio (SNR), which is the difference of these on dB scale, reverberance, etc.

[0026] Other characteristics of the call might also be captured for various purposes, such as detecting causes of errors and customizing error recovery strategies. These characteristics may include, for example, the history of previous errors, the length of the speaker's utterance, the utterance's confidence score, the prompt duration, the amount of time it took the speaker to start

speaking after the prompt, and whether or not the speaker barged-in over the prompt.

[0027] The statistical characterizations of the speaker and acoustics can be made with a statistical classifier. A statistical classifier and a process to train and use it are described below.

[0028] Figure 1 illustrates a network configuration in which a person may communicate with a speech recognition system in accordance with the present invention. The person 1 uses a communication device 2, such as a telephone, to communicate with speech recognition system 3. The communication may take place over a network 4, which may be or may include, for example, the public switched telephone network (PSTN), a wireless telecommunications network, one or more computer networks (e.g., the Internet, a campus intranet, a local area network, a wide area network), or a combination thereof. Note, however, that the speaker 1 could alternatively have a direct audio interface to the system 3, rather than through a network.

[0029] Figure 2 illustrates a process, according to one embodiment, by which the speech recognition system 3 customizes a dialog with the speaker 1, based on one or more characteristics detected automatically by the system 3. At block 201 the system 3 receives in the speech from the speaker 1 during a dialog with the speaker 1. At block 202 the system 3 automatically detects one or more characteristics associated with the dialog, such as any of those mentioned above. Techniques for detecting such characteristics are described in greater detail

8

below. At block 203, the system 3 customizes the dialog for the speaker 1 based on the detected characteristic(s). As mentioned above, the system 3 may also detect and or analyze off-line the characteristics of one or more dialogs, as described further below.

[0030] Figure 3 illustrates in greater detail the speech recognition system 3, according to one embodiment. The system 3 includes a front end 31, a recognition engine 32, a speech application 33, a speech generator 34, a characteristic detector 35, and customization logic 36. The system 3 further includes a set of models 37, content 38 used by the speech application 33, and a set of prompts 39. The system 3 may also include an offline analysis module 40 and an output interface 41, as described below.

[0031] An input speech signal is received by the front end 31 via a microphone, telephony interface, computer network interface, or any other conventional communications interface. The front end 31 is responsible for processing the speech waveform to transform it into a sequence of data points that can be better modeled than the raw waveform. Hence, the front end 31 digitizes the speech waveform (if it is not already digitized), endpoints the speech, and extracts feature vectors from the digitized speech representing spectral components of the speech. In some embodiments, endpointing precedes feature extraction, while in others feature extraction precedes endpointing. The extracted feature vectors are provided to the recognition engine 32, which references the feature vectors against the models 37 to generate recognized speech data. The models 37

9

include a dictionary, acoustic models, and recognition grammars and/or language models.

[0032]   The recognition engine 32 provides the recognized speech to the speech application 33. The speech application 33 is the component of the system which is responsible for performing application-level functions that fulfill the speaker's purpose for contacting the system 3, such as providing stock quotes, booking travel reservations, or performing online banking functions. The speech application 33 may include a natural language interpreter (not shown) to interpret the meaning of the recognized speech. The speech application 33 accesses content 38 in response to the speaker's speech and uses the content 38 to perform its application-level functions. The content 38 may include audio, text, graphics, images, multimedia data, or any combination thereof. When appropriate, the speech application 33 causes content to be output to the speaker by the speech generator 34, by a graphical user interface (not shown), or by any other type of user interface suitable for the content type.

[0033]   The speech generator 34 outputs recorded or synthesized speech to the user, typically over the same communication channel by which the users speech is received by the system 3. The speech generator 34 outputs the speech in response to signals from the speech application 33 and, in some cases, based on the content provided by the speech application 33. The speech generator 34 also responds to the speech application 33 by selecting various prompts 39 to be output to the speaker.

10

[0034] The characteristic detector 35 automatically detects characteristics, such as any of those mentioned above. Depending on the type of characteristic(s) to be detected, the detector 35 may detect the characteristics from the raw input speech, the feature vectors output by the front-end, or both. Detected characteristics are provided to the customization logic 36 and/or are stored in a database 42 for later use (e.g., by the analysis module 40).

[0035] In response to receiving detected characteristics from the detector 35, the customization logic 36 provides control signals to other components of the system 3 to customize the dialog for the speaker at the application level in real-time. To do this, the customization logic 36 may control: the recognition engine 32, the speech application 33, or the speech generator 34, for example. As described further below, examples of application-level customizations that can be performed during a dialog include customized call routing, error recovery, call flow, content selection and delivery, prompt selection and delivery, grammar selection, and persona selection.

[0036] In one embodiment, the speech recognition system 3 includes an analysis module 40 to perform off-line analysis of the stored data representing detected characteristics, and an output interface 41 to output reports of such analyses. These components can be used to provide, among other things, demographic reports with respect to one or more specified characteristics (e.g., callers' age, gender, or nationality), based on large samples of speakers and/or dialogs. Note, however, that the system 3 can be implemented without the

11

analysis module 40 and the output interface 41, as described above. In another

embodiment, the analysis module 40 and the output interface 41 are

implemented separately, i.e., independent from the illustrated speech recognition

system 3, although the analysis module 40 still has access to the data 42

representing detected characteristics in such an embodiment.

[0037]   In yet another embodiment, the characteristic detector 35 is

implemented in a separate system, with the analysis module 40 and the output

interface 41. In such an embodiment, the detection of characteristics is also

performed off-line, based on stored speech data (representing raw speech,

feature vectors, or recognized speech), and analyses and reports are generated

off-line as described above.

[0038]   Figures 4A and 4B illustrate two embodiments of a process for off-line

analysis of dialog data based on detected characteristics, according to one

embodiment. In the embodiment of Figure 4A, the detection of characteristics is

performed in real-time (i.e., during the dialogs), whereas in the embodiment of

Figure 4B the detection of characteristics is performed off-line based on stored

data. Referring first to Figure 4A, at block 401 the system receives input speech

from the speaker. At block 402 the system detects one or more specified

characteristics, such as any of those mentioned above. A technique for detecting

the characteristics is described in greater detail below. The system then stores

data indicative of the detected characteristic(s) at block 403. The foregoing

operations are then repeated for a predetermined number N of dialogs or

12

speakers, or until a report is requested (block 404). When sufficient data is collected or a report is requested, the system analyzes the data with respect to the detected characteristic(s) at block 405 and generates a report of the analysis at block 406. The analysis and report may provide, for example, demographic information with respect to one or more specified characteristics, for a large sample of speakers or dialogs.

[0039]   In the process of Figure 4B, at block 407 the system receives input speech from the speaker. At block 408 the system stores data representative of the input speech (i.e., representing raw speech, feature vectors, or recognized speech). The foregoing operations are repeated for a predetermined number N of dialogs or speakers, or until a report is requested (block 409). At block 410 the system detects one or more specified characteristics based on the stored data. The system analyzes the characteristics data at block 411 and generates a report of the analysis at block 412.

II. Characteristic Detection

[0040]   A technique by which characteristics can be detected will now be described in greater detail. Figure 5 shows an overview of a general process that can be used to train a statistical classifier for any of the characteristics mentioned above. The process is data driven. First, data is collected and manually labeled for each of the conditions of interest (e.g., male callers, female callers) at block 501. A sequence of acoustic features is extracted for each utterance in the data set at block 502. The resulting acoustic features from all utterances are clustered at

block 503 based on similarity. Parameters describing the distributions of these clusters (typically means, variances, and weights for Gaussian mixtures) are estimated from the clustered features at block 504. Collectively, these parameters describe an acoustic model 505 for the corresponding manually labeled condition.

[0041] The acoustic feature vectors include short-time estimates of the spectral components of the signal (e.g., the spectral envelope) and the rate of change and acceleration of these components (temporal derivatives of the short time estimates). They may also include short-time estimates of pitch, the presence of voicing, and the rate of change and acceleration of these.

[0042] In cases where there is a limited set of possible classifications (e.g., only "male" or "female" in the case of speaker's gender), acoustic models can be built from a large collection of example utterances for each classification (e.g., male and female). This leads to classification-specific acoustic models (e.g., male models and female models).

[0043] Figure 6 shows a process by which the above-mentioned acoustic models can be used to make classification decisions, i.e., to detect characteristics. A new (unlabeled) utterance 601 of the speaker is received by the speech recognition system, and a sequence of acoustic features is extracted from the new utterance at block 602. Each classification-specific acoustic model 603 is then evaluated at points described by the sequence of features. The evaluation of the model at each point in the sequence provides an estimate of the probability of

14

observing each feature vector given a specific acoustic model. Total scores for each model are accumulated over the sequence of features by adding the weighted log of the probability of each feature vector. These scores correlate to the probability of the feature sequence for the new utterance given each condition-specific acoustic model (i.e., male or female).

[0044] The total scores are weighted at blocks 604 by the expected likelihood for each condition (e.g., the expected percentage of male and female callers) and, if desired, by additional weights to correct for any asymmetric cost functions (if, for example, the cost of misclassifying a man were higher than misclassifying a woman). After this weighting, the model corresponding to the highest score is selected at block 605 to determine the classification decision 606, i.e., to detect the characteristic.

[0045] In cases where there are many possible classifications (e.g., in the case of classifying the speaker's emotional state), the following approach can be used. Labeled data is collected and acoustic models are built for the classification to be detected (e.g., the emotional state, "frustrated"), and additional, background acoustic models are built from data representing other emotional states not to be explicitly detected. The same process as described with respect to Figure 6 can be used, and the final decision of choosing the model with the highest score now represents a detection decision. If, for example, the system is used to detect frustration of the speaker, and the background model scores higher than the acoustic model trained with speech from frustrated callers, then the decision (the

characteristic detection) is that the caller is not frustrated. When a background modeling approach is used, it may be desirable to require that the detected model score higher than the background model by more than a tunable threshold.

[0046] This same approach can be used to detect essentially any characteristics. For example, to detect gender, male and female models are used; to detect type of speech, models of side speech, background speech, fast and slow speech, accented speech and normal speech are used. Similarly to detect the use of a hands-free telephone vs. a hand-held telephone, or "in a bar" vs. "at an airport," models built from data collected for each condition can be used.

[0047] The above-described approach is most effective when only a small number of classification decisions are to be made (e.g., female vs. male, or hands-free vs. hand-held) for any given utterance. When a large number of classifications are required simultaneously, however (e.g. age, gender, speech type, and environment type), the above approach may require an undesirably large number of data collections and acoustic models. For these applications, instead of building explicit acoustic models for every possible classification, caller-independent transforms can be used to map acoustic models from one channel to generate a synthesized model in a new channel not yet seen. Transforms of this type can be created by using the techniques described in U.S. Patent no. 6,233,556 of Teunen et al., which is incorporated herein by reference.

16

[0048] For example, this separation of channel and speaker allows a set of age-classification models trained from data collected in relatively noiseless ("clean") hand-held environments to be used to detect a caller's age in a more noisy, hands-free environment. In this case, generic transforms for mapping clean handheld acoustic data to hands-free data are applied to the individual age classification models, to generate synthetic age classification algorithms for hands-free environments.

[0049] In one embodiment of a system that makes both channel-type and speaker-related classifications, the channel (e.g., hand-held vs. hands-free) is first identified from models adapted for specific channels. Then, the speaker classification (e.g., age) is identified by choosing from models that are synthesized for that environment by applying the caller-independent transforms. Figure 7 illustrates an example of this process in the case where the channel classification is between "hands-free" and "hand-held", and the age classification is from among the age groups 4-12, 12-20, 20-60, and 60 and above.

[0050] As noted above, the automatic detection of characteristics may also include direct measurements. For example, the log magnitude of the speech level and background noise level, and the difference between these quantities (the SNR), can also be used to customize dialogs. The background noise level can be computed by averaging measurements of the intensity of the input signal over regions that are not detected as speech by the recognition system. Similarly the speech level can be determined by averaging measurements of the input

17

intensity in regions that are detected as speech. On a log scale, the difference between these measurements is the SNR.

[0051] Speech level, background noise level, and SNR can also be used in combination with other measurements, including the history of previous errors, the length of the speaker's utterance, the utterance's confidence score, the prompt duration, start of speech delay and barge-in, to detect why the speaker encountered an error. A statistical classifier as described above could determine the most likely cause of error, although a classifier using another technology, such as neural networks, decision trees or discriminant analysis might alternatively be used.

III. Dialog Customization

[0052] Selected examples of customizing a dialog at the application level will now be described in greater detail. One customization which can be done using this approach is call routing, where the technology routes a call from a person based on the detected characteristic(s). After one or more characteristics of the caller (speaker) and/or his environment are detected, algorithms determine if and where the call should be routed, based on the detected characteristics. For example, if a native Spanish speaker in the Southeastern U.S. calls into a system, the system may detect his Spanish accent. After detection, a system component matches the detected accent with contact information for bilingual, Spanish-speaking operators. The system then forwards the call to the bilingual operators for improved service.

18

[0053]   A second type of dialog customization available is in error recovery

dialogs. In this approach, detected characteristics are used in conjunction with

errors returned from the recognition engine to improve the error recovery

dialogs. If the application detects that an underlying speaker or acoustic

characteristic is relevant to an error that the speaker has just encountered, the

dialog strategy can be customized to help the speaker recover from the error. For

example, if a recognition error is returned and the customization logic detects a

likely cause of the error, such as high background noise, then the dialog can be

adjusted to inform the speaker about the probable cause of the error and possibly

attempt to correct for it. In the case of background noise, the prompt might say

something like, "Sorry, I'm having trouble understanding because there seems to

be a lot of noise. I might have an easier time if you move to a quieter area."

[0054]   A third approach is customizing content, one example of which is

advertising content. In this approach, detected characteristics are used in

conjunction with a dynamic source of content to customize content provided to

the speaker. Algorithms are closely linked with the dynamic content, so

customized content/dialogs can be presented to the speaker. Consider the case

of a young female caller in a drugstore speech application. In this case, the

young female caller may have a stronger interest in particular pharmaceutical

products than an older male caller. Thus, she may want to hear certain dialogs

that contain particular advice and product recommendations, while the older

male caller will want to hear different dialogs. In this case, the system first

19

detects the speaker characteristics (e.g., age and gender) and acoustic characteristics. Algorithms then interact with the detected characteristics to determine if and how the content should be dynamically served in the dialog. The end result is a dialog that is tailored to the young female caller.

[0055] Another approach is to customize prompt delivery, such as to customize speed and/or pausing within and between prompts. In this approach, detected characteristics are used to determine if the prompts should be sped up or slowed down and if pauses should be inserted or shortened. If certain speaker or environmental characteristics are detected that align with preferred prompt speeds and pausing by speaker and environment, prompt speed parameters and pausing algorithms can be adjusted accordingly. For example, an elderly speaker may want to hear slower prompts with more pauses between sentences, than a younger person would prefer. The speaker ultimately hears an adjusted dialog if these parameters are adjusted.

[0056] Yet another approach is to customize the system persona. The persona is the "personality" of the voice output by the speech recognition system. Personas can vary from being very business-like and formal to being very informal and non-mainstream. Certain types of speakers may respond more favorably to certain types of personas. For example, a caller from Texas might respond more favorably to a persona with a Texas accent. The process flow of this approach can be very similar to that of speed and pausing customization.

20

However, rather than modifying speed parameters and pausing algorithms, different prompts are used that reflect a different system persona.

[0057]   Another approach is to customize prompt style. A prompt style is similar to a persona, but it is the wording of the recorded messages and prompts, rather than being a combination of the prompt wording and voice (as in the case of persona). For example, if the system detects that the speaker is using a hands-free device, the system may avoid prompt wording that makes reference to using the touchtone keypad. The process flow of this approach can be very similar to that of speed and pausing customization. However, rather than modifying speed parameters and pausing algorithms, different prompt wording is used. Prompt style can also be used for text-to-speech synthesis. For example, if a speaker from the Southern U.S. is identified, the system might insert slang or other words associated with that region, such as "y'all."

[0058]   Another approach is to customize call flow. Call flow refers to the order in which a caller hears a series of prompts. An older caller may wish to hear slower, more explicit information. For example, if the system detects that the caller is frustrated, then the system might select a call flow embodying a back-off strategy to ensure the caller has better success, or the customization logic might transfer the frustrated caller to a human operator. The process flow of this approach can be very similar to that of content customization. However, rather than content being dynamically served, the flow of the call is dynamically

modified based on expected caller preferences and that match different call flow options to detected characteristics.

[0059] Yet another approach is to customize grammars. Grammars describe the set of words the speech recognition engine is expecting the caller to say at any given point in time. Within this set of expected utterances, certain words and phrases are marked as more probable or not. Grammar customization is important, because speech recognition accuracy depends upon using the most relevant grammar and properly weighting the probabilities of words within these grammars. Dialog customization can be used to dynamically switch grammars for different callers to optimize recognition accuracy. The process flow of this approach can be very similar to that of content customization. However, rather than content being dynamically served, grammars are automatically switched or adjusted based on detected characteristics.

IV. Computer System Implementation

[0060] The speech recognition system described above can be implemented using one or more general-purpose computing platforms, which may be, for example, personal computers (PCs), server-class computers, and/or small form factor computing devices (e.g., cellular telephones and personal digital assistants (PDAs)), or a combination thereof. Figure 8 is a block diagram showing an abstraction of the hardware components of such a computer system. Note that there are many possible implementations represented by this abstraction, which will be readily appreciated by those skilled in the art given this description.

22

[0061]   The illustrated system includes one or more processors 81 (i.e. a central processing unit (CPU)), read-only memory (ROM) 82, and random access memory (RAM) 83, which may be coupled to each other by a bus system 88 and/or by direct connections. The processor(s) 81 may be, or may include, one or more programmable general-purpose or special-purpose microprocessors, digital signal processors (DSPs), programmable controllers, application specific integrated circuits (ASICs), programmable logic devices (PLDs), or a combination of such devices. The bus system (if any) 88 includes one or more buses or other connections, which may be connected to each other through various bridges, controllers and/or adapters, such as are well-known in the art. For example, the bus system 88 may include a "system bus", which may be connected through one or more adapters to one or more expansion buses, such as a Peripheral Component Interconnect (PCI) bus, HyperTransport or industry standard architecture (ISA) bus, small computer system interface (SCSI) bus, universal serial bus (USB), or Institute of Electrical and Electronics Engineers (IEEE) standard 1394 bus (sometimes referred to as "Firewire").

[0062]   Also coupled to the bus system 88 are one or more mass storage devices 84, an audio processing module 85, a data communication device 87, and one or more other input/output (I/O) devices 86. Each mass storage device 84 may be, or may include, any one or more devices suitable for storing large volumes of data in a non-volatile manner, such as a magnetic disk or tape, magneto-optical (MO) storage device, or any of various forms of Digital Versatile Disk (DVD) or

CD-based storage, or a combination thereof.

[0063] The data communication device 87 is one or more data communication devices suitable for enabling the processing system to communicate data with remote devices and systems via an external communication link 90. Each such data communication device may be, for example, an Ethernet adapter, a Digital Subscriber Line (DSL) modem, a cable modem, an Integrated Services Digital Network (ISDN) adapter, a satellite transceiver, or the like. Other I/O devices 86 may be included in some embodiments and may include, for example, a keyboard or keypad, a display device, and a pointing device (e.g., a mouse, trackball, or touchpad).

[0064] Of course, certain components shown in Figure 8 may not be needed for certain embodiments. For example, a data communication device may not be needed where the ASR system is embedded in a single processing device. Similarly, a keyboard and display device may not be needed in a device which operates only as a server.

[0065] The above-described processes and techniques for characteristic detection, dialog customization, and data analysis may be implemented at least partially in software, which may be stored and/or executed on a computer system such as shown in Figure 8. For example, such software may be software 89 residing, either entirely or in part, in any of ROM 82, RAM 83, or mass storage device(s) 84. Such software may be executed by the processor(s) 81 to carry out the above-described processes. Alternatively, the above-described processes and

24

techniques can be implemented using hardwired circuitry (e.g., ASICs, PLDs), or using a combination of hardwired circuitry and circuitry programmed with software.

[0066]   Thus, a method and apparatus for customizing a human-machine dialog or analyzing data relating to human-machine dialogs, based on detection of characteristics associated with one or more dialogs, have been described. Although the present invention has been described with reference to specific exemplary embodiments, it will be evident that various modifications and changes may be made to these embodiments without departing from the broader spirit and scope of the invention as set forth in the claims. Accordingly, the specification and drawings are to be regarded in an illustrative sense rather than a restrictive sense.